On the φ-family of probability distributions[☆]

Rui F. Vigelis*, Charles C. Cavalcante

Wireless Telecommunication Research Group, Department of Teleinformatics Engineering, Federal University of Ceará, Fortaleza-CE, Brazil

Abstract

We generalize the exponential family of probability distributions $\mathcal{E}_p$. In our approach, the exponential function is replaced by the φ-function, resulting in the φ-family of probability distributions $\mathcal{F}_c^\varphi$. We provide how φ-families are constructed. In the φ-family, the analogous of the cumulant-generating functional is a normalizing function. We define the φ-divergence as the Bregman divergence associated to the normalizing function, providing a generalization of the Kullback–Leibler divergence. We found that the Kaniadakis' κ-exponential function satisfies the definition of φ-functions. A formula for the φ-divergence where the φ-function is the κ-exponential function is derived.

Keywords: Exponential family of probability distributions, Musielak–Orlicz spaces, Bregman divergence

1. Introduction

Let (T, Σ, μ) be a σ-finite, non-atomic measure space. We denote by $\mathcal{P}_\mu = \mathcal{P}(T, \Sigma, \mu)$ the family of all probability measures on T that are equivalent to the measure μ. The probability family $\mathcal{P}_\mu$ can be represented as (we adopt the same symbol $\mathcal{P}_\mu$ for this representation)

$$\mathcal{P}_\mu = \{p \in L^0 : p > 0 \text{ and } \mathbb{E}[p] = 1\},$$

where L^0 is the linear space of all real-valued, measurable functions on T, with equality μ-a.e., and $\mathbb{E}[\cdot]$ denotes the expectation with respect to the measure μ.

The family $\mathcal{P}_\mu$ can be equipped with a structure of C^∞-Banach manifold, using the Orlicz space $L^{\Phi_1}(p) = L^{\Phi_1}(T, \Sigma, p \cdot \mu)$ associated to the Orlicz function $\Phi_1(u) = \exp(u) - 1$, for $u \geq 0$. With this structure, $\mathcal{P}_\mu$ is called the *exponential*

[☆]This work received financial support from CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

*Corresponding author

Email addresses: `rfvigelis@gtel.ufc.br` (Rui F. Vigelis), `charles@gtel.ufc.br` (Charles C. Cavalcante)

statistical manifold., whose construction was proposed in [1] and developed in [2, 3, 4]. Each connected component of the exponential statistical manifold gives rise to an *exponential family of probability distributions* $\mathcal{E}_p$ (for each $p \in \mathcal{P}_\mu$). Each element of $\mathcal{E}_p$ can be expressed as

$$\boldsymbol{e}_p(u) = e^{u - K_p(u)} p, \qquad \text{for } u \in \mathcal{B}_p, \tag{1}$$

for a subset $\mathcal{B}_p$ of the Orlicz space $L^{\Phi_1}(p)$. K_p is the cumulant-generating functional $K_p(u) = \log \mathbb{E}_p[e^u]$, where $\mathbb{E}_p[\cdot]$ is the expectation with respect to $p \cdot \mu$. If c is a measurable function such that $p = e^c$, then (1) can be rewritten as

$$\boldsymbol{e}_p(u) = e^{c + u - K_p(u) \cdot \mathbf{1}_T}, \qquad \text{for } u \in \mathcal{B}_p, \tag{2}$$

where $\mathbf{1}_A$ is the indicator function of a subset $A \subseteq T$.

In the *φ-family of probability distributions* $\mathcal{F}_c^\varphi$, which we propose, the exponential function is replaced by the so called *φ-function* $\varphi\colon T \times \overline{\mathbb{R}} \to [0, \infty]$. The function $\varphi(t, \cdot)$ has a "shape" which is similar to that of an exponential function, with an arbitrary rate of increasing. For example, we found that the κ-exponential function satisfies the definition of φ-functions. As in the exponential family, the φ-families are the connected component of $\mathcal{P}_\mu$, which is endowed with a structure of C^∞-Banach manifold, using φ in the place of an exponential function. Let c be any measurable function such that $\varphi(t, c(t))$ belongs to $\mathcal{P}_\mu$. The elements of the φ-family of probability distributions $\mathcal{F}_c^\varphi$ are given by

$$\boldsymbol{\varphi}_c(u)(t) = \varphi(t, c(t) + u(t) - \psi(u) u_0(t)), \qquad \text{for } u \in \mathcal{B}_c^\varphi, \tag{3}$$

for a subset $\mathcal{B}_c^\varphi$ of a Musielak–Orlicz space L_c^φ. The *normalizing function* $\psi\colon \mathcal{B}_c^\varphi \to [0, \infty)$ and the measurable function $u_0\colon T \to [0, \infty)$ in (3) replaces K_p and $\mathbf{1}_T$ in (2), receptively. The function u_0 is not arbitrary. In the text, we will show how u_0 can be chosen.

We define the *φ-divergence* as the a Bregman divergence associated to the normalizing function ψ, providing a generalization of the Kullback–Leibler divergence. Then geometrical aspects related to the φ-family can be developed, since the Fisher information (from which the Information Geometry [5, 6] is based) is derived from the divergence. A formula for the φ-divergence where the φ-function is the κ-exponential function is derived, which we called the *κ-divergence*.

We expect that an extension of our work will provide advances in other areas, like in Information Geometry or in the non-parametric, non-commutative setting [7, 8]. The rest of this paper is organized as follows. Section 2 deals with the topics of Musielak–Orlicz spaces we will use in the the construction of the φ-family of probability distributions. In Section 3, the exponential statistical manifold is reviewed. The construction of the φ-family of probability distributions is given in Section 4. Finally, the φ-divergence is derived in Section 5.

2. Musielak–Orlicz spaces

In this section we provide a brief introduction to Musielak–Orlicz (function) spaces, which are used in the construction of the exponential and φ-families. A more detailed exposition about these spaces can be found in [9, 10, 11].

We say that $\Phi\colon T \times [0,\infty] \to [0,\infty]$ is a *Musielak–Orlicz function* when, for μ-a.e. $t \in T$,

(i) $\Phi(t,\cdot)$ is convex and lower semi-continuous,

(ii) $\Phi(t,0) = \lim_{u\downarrow 0} \Phi(t,u) = 0$ and $\Phi(t,\infty) = \infty$,

(iii) $\Phi(\cdot,u)$ is measurable for all $u \geq 0$.

Items (i)–(ii) guarantee that $\Phi(t,\cdot)$ is not equal to 0 or ∞ on the interval $(0,\infty)$. A Musielak–Orlicz function Φ is said to be an *Orlicz function* if the functions $\Phi(t,\cdot)$ are identical for μ-a.e. $t \in T$.

Define the functional $I_\Phi(u) = \int_T \Phi(t,|u(t)|)d\mu$, for any $u \in L^0$. The *Musielak–Orlicz space*, *Musielak–Orlicz class*, and *Morse–Transue space*, are given by

$$
\begin{aligned}
L^\Phi &= \{u \in L^0 : I_\Phi(\lambda u) < \infty \text{ for some } \lambda > 0\},\\
\tilde{L}^\Phi &= \{u \in L^0 : I_\Phi(u) < \infty\},
\end{aligned}
$$

and

$$
E^\Phi = \{u \in L^0 : I_\Phi(\lambda u) < \infty \text{ for all } \lambda > 0\},
$$

respectively. If the underlying measure space (T,Σ,μ) have to be specified, we write $L^\Phi(T,\Sigma,\mu)$, $\tilde{L}^\Phi(T,\Sigma,\mu)$ and $E^\Phi(T,\Sigma,\mu)$ in the place of L^Φ, $\tilde{L}^\Phi$ and E^Φ, respectively. Clearly, $E^\Phi \subseteq \tilde{L}^\Phi \subseteq L^\Phi$. The Musielak–Orlicz space L^Φ can be interpreted as the smallest vector subspace of L^0 that contains $\tilde{L}^\Phi$, and E^Φ is the largest vector subspace of L^0 that is contained in $\tilde{L}^\Phi$.

The Musielak–Orlicz space L^Φ is a Banach space when it is endowed with the *Luxemburg norm*

$$
\|u\|_\Phi = \inf\Big\{\lambda > 0 : I_\Phi\Big(\frac{u}{\lambda}\Big) \leq 1\Big\},
$$

or the *Orlicz norm*

$$
\|u\|_{\Phi,0} = \sup\bigg\{\bigg|\int_T uv d\mu\bigg| : v \in \tilde{L}^{\Phi^*} \text{ and } I_{\Phi^*}(v) \leq 1\bigg\},
$$

where $\Phi^*(t,v) = \sup_{u\geq 0}(uv - \Phi(t,u))$ is the *Fenchel conjugate* of $\Phi(t,\cdot)$. These norms are equivalent and the inequalities $\|u\|_\Phi \leq \|u\|_{\Phi,0} \leq 2\|u\|_\Phi$ hold for all $u \in L^\Phi$.

If we can find a non-negative function $f \in \tilde{L}^\Phi$ and a constant $K > 0$ such that

$$
\Phi(t,2u) \leq K\Phi(t,u), \quad \text{for all } u \geq f(t),
$$

then we say that Φ satisfies the Δ_2-*condition*, or belong to the Δ_2-*class* (denoted by $\Phi \in \Delta_2$). When the Musielak–Orlicz function Φ satisfies the Δ_2-condition, E^Φ coincides with L^Φ. On the other hand, if Φ is finite-valued and does not satisfy the Δ_2-condition, then the Musielak–Orlicz class $\tilde{L}^\Phi$ is not open and its interior coincides with

$$B_0(E^\Phi, 1) = \{u \in L^\Phi : \inf_{v \in E^\Phi} \|u - v\|_{\Phi,0} < 1\},$$

or, equivalently, $B_0(E^\Phi, 1) \subsetneq \tilde{L}^\Phi \subsetneq \overline{B}_0(E^\Phi, 1)$.

3. The exponential statistical manifold

This section starts with the definition of a C^k-Banach manifold. A C^k-*Banach manifold* is a set M and a collection of pairs $(U_\alpha, \boldsymbol{x}_\alpha)$ (α belonging to some indexing set), composed by open subsets U_α of some Banach space X_α, and injective mappings $\boldsymbol{x}_\alpha \colon U_\alpha \to M$, satisfying the following conditions:

(bm1) the sets $\boldsymbol{x}_\alpha(U_\alpha)$ cover M, i.e., $\bigcup_\alpha \boldsymbol{x}_\alpha(U_\alpha) = M$;

(bm2) for any pair of indices α, β such that $\boldsymbol{x}_\alpha(U_\alpha) \cap \boldsymbol{x}_\beta(U_\beta) = W \neq \emptyset$, the sets $\boldsymbol{x}_\alpha^{-1}(W)$ and $\boldsymbol{x}_\beta^{-1}(W)$ are open in X_α and X_β, respectively; and

(bm3) the *transition map* $\boldsymbol{x}_\beta^{-1} \circ \boldsymbol{x}_\alpha \colon \boldsymbol{x}_\alpha^{-1}(W) \to \boldsymbol{x}_\beta^{-1}(W)$ is a C^k-isomorphism.

The pair $(U_\alpha, \boldsymbol{x}_\alpha)$ with $p \in \boldsymbol{x}_\alpha(U_\alpha)$ is called a *parametrization* (or *system of coordinates*) of M at p; and $\boldsymbol{x}_\alpha(U_\alpha)$ is said to be a *coordinate neighborhood* at p.

The set M can be endowed with a topology in a unique way such that each $\boldsymbol{x}_\alpha(U_\alpha)$ is open, and the $\boldsymbol{x}_\alpha$'s are topological isomorphisms. We note that if $k \geq 1$ and two parametrizations $(U_\alpha, \boldsymbol{x}_\alpha)$ and $(U_\beta, \boldsymbol{x}_\beta)$ are such that $\boldsymbol{x}_\alpha(U_\alpha)$ and $\boldsymbol{x}_\beta(U_\beta)$ have a non-empty intersection, then from the derivative of $\boldsymbol{x}_\beta^{-1} \circ \boldsymbol{x}_\alpha$ we have that X_α and X_β are isomorphic.

Two collections $\{(U_\alpha, \boldsymbol{x}_\alpha)\}$ and $\{(V_\beta, \boldsymbol{x}_\beta)\}$ satisfying (bm1)–(bm3) are said to be C^k-*compatible* if their union also satisfies (bm1)–(bm3). It can be verified that the relation of C^k-compatibility is an equivalence relation. An equivalence class of C^k-compatible collections $\{(U_\alpha, \boldsymbol{x}_\alpha)\}$ on M is said to define a C^k-*differentiable structure* on X.

Now we review the construction of the exponential statistical manifold. We consider the Musielak–Orlicz space $L^{\Phi_1}(p) = L^{\Phi_1}(T, \Sigma, p \cdot \mu)$, where the Orlicz function $\Phi_1 \colon [0, \infty) \to [0, \infty)$ is given by $\Phi_1(u) = e^u - 1$, and p is a probability density in $\mathcal{P}_\mu$. The space $L^{\Phi_1}(p)$ corresponds to the set of all functions $u \in L^0$ whose *moment-generating function* $\widehat{u}_p(t) = \mathbb{E}_p[e^{tu}]$ is finite in a neighborhood of 0.

For every function $u \in L^0$ we define the *moment-generating functional*

$$M_p(u) = \mathbb{E}_p[e^u],$$

and the *cumulant-generating functional*

$$K_p(u) = \log M_p(u).$$

Clearly, these functionals are not expected to be finite for every $u \in L^0$. Denote by $\mathcal{K}_p$ the interior of the set of all functions $u \in L^{\Phi_1}(p)$ whose moment-generating functional $M_p(u)$ is finite. Equivalently, a function $u \in L^{\Phi_1}(p)$ belongs to $\mathcal{K}_p$ if and only if $M_p(\lambda u)$ is finite for every λ in some neighborhood of $[0, 1]$. The closed subspace of *p-centered* random variables

$$B_p = \{u \in L^{\Phi_1}(p) : \mathbb{E}_p[u] = 0\}$$

is taken to be the coordinate Banach space. The *exponential parametrization* $\boldsymbol{e}_p \colon \mathcal{B}_p \to \mathcal{E}_p$ maps $\mathcal{B}_p = B_p \cap \mathcal{K}_p$ to the *exponential family* $\mathcal{E}_p = \boldsymbol{e}_p(\mathcal{B}_p) \subseteq \mathcal{P}_\mu$, according to

$$\boldsymbol{e}_p(u) = e^{u - K_p(u)} p, \quad \text{for all } u \in \mathcal{B}_p.$$

$\boldsymbol{e}_p$ is a bijection from $\mathcal{B}_p$ to its image $\mathcal{E}_p = \boldsymbol{e}_p(\mathcal{B}_p)$, whose inverse $\boldsymbol{e}_p^{-1} \colon \mathcal{E}_p \to \mathcal{B}_p$ can be expressed as

$$\boldsymbol{e}_p^{-1}(q) = \log\Big(\frac{q}{p}\Big) - \mathbb{E}_p\Big[\log\Big(\frac{q}{p}\Big)\Big], \quad \text{for } q \in \mathcal{E}_p.$$

Since $K_p(u) < \infty$ for every $u \in \mathcal{K}_p$, we have that $\boldsymbol{e}_p$ can be extended to $\mathcal{K}_p$. The restriction of $\boldsymbol{e}_p$ to $\mathcal{B}_p$ guarantees that $\boldsymbol{e}_p$ is bijective.

Given two probability densities p and q in the same connected component of $\mathcal{P}_\mu$, the exponential probability families $\mathcal{E}_p$ and $\mathcal{E}_q$ coincide, and the exponential spaces $L^{\Phi_1}(p)$ and $L^{\Phi_1}(q)$ are isomorphic (see [2, Proposition 5]). Hence, $\mathcal{B}_p = \boldsymbol{e}_p^{-1}(\mathcal{E}_p \cap \mathcal{E}_q)$ and $\mathcal{B}_q = \boldsymbol{e}_q^{-1}(\mathcal{E}_p \cap \mathcal{E}_q)$. The transition map $\boldsymbol{e}_q^{-1} \circ \boldsymbol{e}_p : \mathcal{B}_p \to \mathcal{B}_q$, which can be written as

$$\boldsymbol{e}_q^{-1} \circ \boldsymbol{e}_p(u) = u + \log\Big(\frac{p}{q}\Big) - \mathbb{E}_q\Big[u + \log\Big(\frac{p}{q}\Big)\Big], \quad \text{for all } u \in \mathcal{B}_p,$$

is a C^∞-function. Clearly, $\bigcup_{p \in \mathcal{P}_\mu} e_p(\mathcal{B}_p) = \mathcal{P}_\mu$. Thus the collection $\{(\mathcal{B}_p, \boldsymbol{e}_p)\}_{p \in \mathcal{P}_\mu}$ satisfies (bm1)–(bm2). Hence $\mathcal{P}_\mu$ is a C^∞-Banach manifold, which is called the *exponential statistical manifold*.

4. Construction of the φ-family of probability distributions

The generalization of the exponential family is based on the replacement of the exponential function by a *φ-function* $\varphi \colon T \times \overline{\mathbb{R}} \to [0, \infty]$ that satisfies the following properties, for μ-a.e. $t \in T$:

(a1) $\varphi(t, \cdot)$ is convex and injective,

(a2) $\varphi(t, -\infty) = 0$ and $\varphi(t, \infty) = \infty$,

(a3) $\varphi(\cdot, u)$ is measurable for all $u \in \mathbb{R}$.

In addition, we assume a positive, measurable function $u_0\colon T \to (0,\infty)$ can be found such that, for every measurable function $c\colon T \to \mathbb{R}$ for which $\varphi(t, c(t))$ is in $\mathcal{P}_\mu$, we have that

(a4) $\varphi(t, c(t) + \lambda u_0(t))$ is μ-integrable for all $\lambda > 0$.

The choice for $\varphi(t,\cdot)$ injective with image $[0,\infty]$ is justified by the fact that a parametrization of $\mathcal{P}_\mu$ maps real-valued functions to positive functions. Moreover, by (a1), $\varphi(t,\cdot)$ is continuous and strictly increasing. From (a3), the function $\varphi(t, u(t))$ is measurable if and only if $u\colon T \to \mathbb{R}$ is measurable. Replacing $\varphi(t,u)$ by $\varphi(t, u_0(t)u)$, a "new" function $u_0 = 1$ is obtained satisfying (a4).

Example 1 ([12], [13], [14])**.** The *Kaniadakis' κ-exponential* $\exp_\kappa\colon \mathbb{R} \to (0,\infty)$ for $\kappa \in [-1,1]$ is defined as

$$\exp_\kappa(u) = \begin{cases} (\kappa u + \sqrt{1+\kappa^2 u^2})^{1/\kappa}, & \text{if } \kappa \neq 0, \\ \exp(u), & \text{if } \kappa = 0. \end{cases}$$

The inverse of $\exp_\kappa$ is the *Kaniadakis' κ-logarithm*

$$\ln_\kappa(u) = \begin{cases} \dfrac{u^\kappa - u^{-\kappa}}{2\kappa}, & \text{if } \kappa \neq 0, \\ \ln(u), & \text{if } \kappa = 0. \end{cases}$$

Some algebraic properties of the ordinary exponential and logarithm functions are preserved:

$$\exp_\kappa(u)\exp_\kappa(-u) = 1, \qquad \ln_\kappa(u) + \ln_\kappa(u^{-1}) = 0.$$

For a measurable function $\kappa\colon T \to [-1,1]$, we define the *variable κ-exponential* $\exp_\kappa\colon T \times \mathbb{R} \to (0,\infty)$ as

$$\exp_\kappa(t,u) = \exp_{\kappa(t)}(u),$$

whose inverse is called the *variable κ-logarithm*:

$$\ln_\kappa(t,u) = \ln_{\kappa(t)}(u).$$

Assuming that $\kappa_- = \operatorname{ess\,inf}|\kappa(t)| > 0$, the variable κ-exponential $\exp_\kappa$ satisfies (a1)–(a4). The verification of (a1)–(a3) is easy. Moreover, we notice that $\exp_\kappa(t,\cdot)$ is strictly convex. We can write for $\alpha \geq 1$

$$\begin{aligned} \exp_\kappa(t,\alpha u) &= (\kappa(t)\alpha u + \alpha\sqrt{1/\alpha^2 + \kappa(t)^2 u^2})^{1/\kappa} \\ &\leq \alpha^{1/|\kappa|}(\kappa(t)u + \sqrt{1+\kappa(t)^2 u^2})^{1/\kappa} \\ &\leq \alpha^{1/\kappa_-}\exp_\kappa(t,u). \end{aligned}$$

By the convexity of $\exp_\kappa(t,\cdot)$, we obtain for any $\lambda \in (0,1)$

$$\begin{aligned} \exp_\kappa(t, c+u) &\leq \lambda \exp_\kappa(t, \lambda^{-1}c) + (1-\lambda)\exp_\kappa(t,(1-\lambda)^{-1}u) \\ &\leq \lambda^{1-1/\kappa_-}\exp_\kappa(t,c) + (1-\lambda)^{1-1/\kappa_-}\exp_\kappa(t,u). \end{aligned}$$

Thus any positive function u_0 such that $\mathbb{E}[\exp_\kappa(u_0)] < \infty$ satisfies (a4).

Let $c\colon T \to \mathbb{R}$ be a measurable function such that $\varphi(t, c(t))$ is μ-integrable. We define the Musielak–Orlicz function

$$\Phi(t, u) = \varphi(t, c(t) + u) - \varphi(t, c(t)).$$

and denote L^Φ, $\tilde{L}^\Phi$ and E^Φ by L_c^φ, $\tilde{L}_c^\varphi$ and E_c^φ, respectively. Since $\varphi(t, c(t))$ is μ-integrable, the Musielak–Orlicz space L_c^φ corresponds to the set of all functions $u \in L^0$ for which $\varphi(t, c(t) + \lambda u(t))$ is μ-integrable for every λ contained in some neighborhood of 0. By the convexity of $\varphi(t, \cdot)$, we have

$$u\varphi'(t, c(t)) \leq \varphi(t, c(t) + u) - \varphi(t, c(t)), \text{ for all } u \in \mathbb{R}. \tag{4}$$

Hence every function u in L_c^φ belongs to the weighted Lebesgue space $L_w^1(\mu)$ where $w(t) = \varphi'(t, c(t))$.

Let $\mathcal{K}_c^\varphi$ be the set of all functions $u \in L_c^\varphi$ such that $\varphi(t, c(t) + \lambda u(t))$ is μ-integrable for every λ in a neighborhood of $[0, 1]$. Denote by $\boldsymbol{\varphi}$ the operator acting on the set of real-valued functions $u\colon T \to \mathbb{R}$ given by $\boldsymbol{\varphi}(u)(t) = \varphi(t, u(t))$. For each probability density $p \in \mathcal{P}_\mu$, we can take a measurable function $c\colon T \to \mathbb{R}$ such that $p = \boldsymbol{\varphi}(c)$. The first import result in the construction of the φ-family is given below.

Lemma 2. *The set $\mathcal{K}_c^\varphi$ is open in L_c^φ.*

Proof. Take any $u \in \mathcal{K}_c^\varphi$. We can find $\varepsilon \in (0, 1)$ such that $\mathbb{E}[\boldsymbol{\varphi}(c + \alpha u)] < \infty$ for every $\alpha \in [-\varepsilon, 1 + \varepsilon]$. Let $\delta = [\frac{2}{\varepsilon}(1 + \varepsilon)(1 + \frac{\varepsilon}{2})]^{-1}$. For any function $v \in L_c^\varphi$ in the open ball $B_\delta = \{w \in L_c^\varphi : \|w\|_\Phi < \delta\}$, we have $I_\Phi(\frac{v}{\delta}) \leq 1$. Thus $\mathbb{E}[\boldsymbol{\varphi}(c + \frac{1}{\delta}|v|)] \leq 2$. Taking any $\alpha \in (0, 1 + \frac{\varepsilon}{2})$, we denote $\lambda = \frac{\alpha}{1+\varepsilon}$. In virtue of

$$\frac{\alpha}{1 - \lambda} = \frac{\alpha}{1 - \frac{\alpha}{1+\varepsilon}} \leq \frac{1 + \frac{\varepsilon}{2}}{1 - \frac{1+\frac{\varepsilon}{2}}{1+\varepsilon}} = \frac{2}{\varepsilon}(1 + \varepsilon)(1 + \frac{\varepsilon}{2}) = \frac{1}{\delta},$$

it follows that

$$\begin{aligned}
\boldsymbol{\varphi}(c + \alpha(u + v)) &= \boldsymbol{\varphi}(\lambda(c + \tfrac{\alpha}{\lambda}u) + (1 - \lambda)(c + \tfrac{\alpha}{1-\lambda}v)) \\
&\leq \lambda\boldsymbol{\varphi}(c + \tfrac{\alpha}{\lambda}u) + (1 - \lambda)\boldsymbol{\varphi}(c + \tfrac{\alpha}{1-\lambda}v) \\
&\leq \lambda\boldsymbol{\varphi}(c + (1 + \varepsilon)u) + (1 - \lambda)\boldsymbol{\varphi}(c + \tfrac{1}{\delta}|v|).
\end{aligned} \tag{5}$$

For $\alpha \in (-\frac{\varepsilon}{2}, 0)$, we can write

$$\begin{aligned}
\boldsymbol{\varphi}(c + \alpha(u + v)) &\leq \tfrac{1}{2}\boldsymbol{\varphi}(c + 2\alpha u) + \tfrac{1}{2}\boldsymbol{\varphi}(c + 2\alpha v) \\
&\leq \tfrac{1}{2}\boldsymbol{\varphi}(c + 2\alpha u) + \tfrac{1}{2}\boldsymbol{\varphi}(c + |v|).
\end{aligned} \tag{6}$$

By (5) and (6), we get $\mathbb{E}[\boldsymbol{\varphi}(c + \alpha(u + v))] < \infty$, for any $\alpha \in (-\frac{\varepsilon}{2}, 1 + \frac{\varepsilon}{2})$. Hence the ball of radius δ centered at u is contained in $\mathcal{K}_c^\varphi$. Therefore, the set $\mathcal{K}_c^\varphi$ is open. $\square$

Clearly, for $u \in \mathcal{K}_c^\varphi$ the function $\boldsymbol{\varphi}(c+u)$ is not necessarily in $\mathcal{P}_\mu$. The *normalizing function* $\psi\colon \mathcal{K}_c^\varphi \to \mathbb{R}$ is introduced in order to make the density

$$\boldsymbol{\varphi}(c+u-\psi(u)u_0)$$

contained in $\mathcal{P}_\mu$, for any $u \in \mathcal{K}_c^\varphi$. We have to find the functions for which the normalizing function there exists. For a function $u \in L_c^\varphi$, suppose that $\boldsymbol{\varphi}(c+u-\alpha u_0)$ is μ-integrable for some $\alpha \in \mathbb{R}$. Then u is in the closure of the set $\mathcal{K}_c^\varphi$. Indeed, for any $\lambda \in (0,1)$,

$$\begin{aligned}\boldsymbol{\varphi}(c+\lambda u) &= \boldsymbol{\varphi}(\lambda(c+u-\alpha u_0)+(1-\lambda)(c+\tfrac{\lambda}{1-\lambda}\alpha u_0)) \\ &\le \lambda\boldsymbol{\varphi}(c+u-\alpha u_0)+(1-\lambda)\boldsymbol{\varphi}(c+\tfrac{\lambda}{1-\lambda}\alpha u_0).\end{aligned}$$

Since the function u_0 satisfies (a4), we obtain that $\boldsymbol{\varphi}(c+\lambda u)$ is μ-integrable. Hence the maximal, open domain of ψ is contained in $\mathcal{K}_c^\varphi$.

Proposition 3. *If the function u is in $\mathcal{K}_c^\varphi$, then there exists a unique $\psi(u) \in \mathbb{R}$ for which $\boldsymbol{\varphi}(c+u-\psi(u)u_0)$ is a probability density in $\mathcal{P}_\mu$.*

Proof. We will show that if the function u is in $\mathcal{K}_c^\varphi$, then $\boldsymbol{\varphi}(c+u+\alpha u_0)$ is μ-integrable for every $\alpha \in \mathbb{R}$. Since u is in $\mathcal{K}_c^\varphi$, we can find $\varepsilon > 0$ such that $\boldsymbol{\varphi}(c+(1+\varepsilon)u)$ is μ-integrable. Taking $\lambda = \frac{1}{1+\varepsilon}$, we can write

$$\begin{aligned}\boldsymbol{\varphi}(c+u+\alpha u_0) &= \boldsymbol{\varphi}(\lambda(c+\tfrac{1}{\lambda}u)+(1-\lambda)(c+\tfrac{1}{1-\lambda}\alpha u_0)) \\ &\le \lambda\boldsymbol{\varphi}(c+\tfrac{1}{\lambda}u)+(1-\lambda)\boldsymbol{\varphi}(c+\tfrac{1}{1-\lambda}\alpha u_0).\end{aligned}$$

Thus $\boldsymbol{\varphi}(c+u+\alpha u_0)$ is μ-integrable. By the Dominated Convergence Theorem, the map $\alpha \mapsto J(\alpha) = \mathbb{E}[\boldsymbol{\varphi}(c+u+\alpha u_0)]$ is continuous, tends to 0 as $\alpha \to -\infty$, and goes to infinity as $\alpha \to \infty$. Since $\varphi(t,\cdot)$ is strictly increasing, it follows that $J(\alpha)$ is also strictly increasing. Therefore, there exists a unique $\psi(u) \in \mathbb{R}$ for which $\boldsymbol{\varphi}(c+u-\psi(u)u_0)$ is a probability density in $\mathcal{P}_\mu$. $\square$

The function $\psi\colon \mathcal{K}_c^\varphi \to \mathbb{R}$ can take both positive and negative values. However, if the domain of ψ is restricted to a subspace of L_c^φ, its image will be contained in $[0,\infty)$. Denote the closed subspace

$$B_c^\varphi = \{u \in L_c^\varphi : \mathbb{E}[u\boldsymbol{\varphi}'(c)] = 0\},$$

and let $\mathcal{B}_c^\varphi = B_c^\varphi \cap \mathcal{K}_c^\varphi$. Supposing that $u \in \mathcal{B}_c^\varphi$, it follows that $\mathbb{E}[u\boldsymbol{\varphi}'(c)] = 0$ and $\mathbb{E}[\boldsymbol{\varphi}(c+u)] < \infty$; and, according to inequality (4), we have

$$1 = \mathbb{E}[u\boldsymbol{\varphi}'(c)] + \mathbb{E}[\boldsymbol{\varphi}(c)] \le \mathbb{E}[\boldsymbol{\varphi}(c+u)] < \infty.$$

If $u \in \mathcal{K}_c^\varphi$ belongs to the subspace B_c^φ, the integral of $\boldsymbol{\varphi}(c+u)$ is greater than or equal to 1. Subtracting $\psi(u)u_0$, the integral decreases to 1, and we obtain that $\boldsymbol{\varphi}(c+u-\psi(u)u_0)$ is in $\mathcal{P}_\mu$.

For each measurable function $c\colon T \to \mathbb{R}$ such that the probability density $p = \boldsymbol{\varphi}(c)$ belongs to $\mathcal{P}_\mu$, we associate a parametrization $\boldsymbol{\varphi}_c\colon \mathcal{B}_c^\varphi \to \mathcal{F}_c^\varphi$ that

maps each function u in $\mathcal{B}_c^\varphi$ to a probability density in $\mathcal{F}_c^\varphi = \varphi_c(\mathcal{B}_c^\varphi) \subseteq \mathcal{P}_\mu$ according to

$$\boldsymbol{\varphi}_c(u) = \boldsymbol{\varphi}(c + u - \psi(u)u_0).$$

Clearly, we have $\mathcal{P}_\mu = \bigcup\{\mathcal{F}_c^\varphi : \boldsymbol{\varphi}(c) \in \mathcal{P}_\mu\}$. Moreover, the map $\boldsymbol{\varphi}_c$ is a bijection from $\mathcal{B}_c^\varphi$ to $\mathcal{F}_c^\varphi$. If the functions $u, v \in \mathcal{B}_c^\varphi$ are such that $\boldsymbol{\varphi}_c(u) = \boldsymbol{\varphi}_c(v)$, then the difference $u - v = (\psi(u) - \psi(v))u_0$ is in B_c^φ. Consequently, $\psi(u) = \psi(v)$ and then $u = v$.

Suppose that the measurable functions $c_1, c_2 \colon T \to \mathbb{R}$ are such that $p_1 = \boldsymbol{\varphi}(c_1)$ and $p_2 = \boldsymbol{\varphi}(c_2)$ belong to $\mathcal{P}_\mu$. The parametrizations $\boldsymbol{\varphi}_{c_1} : \mathcal{B}_{c_1}^\varphi \to \mathcal{F}_{c_1}^\varphi$ and $\boldsymbol{\varphi}_{c_2} \colon \mathcal{B}_{c_2}^\varphi \to \mathcal{F}_{c_2}^\varphi$ related to these functions have transition map

$$\boldsymbol{\varphi}_{c_2}^{-1} \circ \boldsymbol{\varphi}_{c_1} : \boldsymbol{\varphi}_{c_1}^{-1}(\mathcal{F}_{c_1}^\varphi \cap \mathcal{F}_{c_2}^\varphi) \to \boldsymbol{\varphi}_{c_2}^{-1}(\mathcal{F}_{c_1}^\varphi \cap \mathcal{F}_{c_2}^\varphi).$$

Let $\psi_1 \colon \mathcal{B}_{c_1}^\varphi \to \mathbb{R}$ and $\psi_2 \colon \mathcal{B}_{c_2}^\varphi \to \mathbb{R}$ be the normalizing functions associated to c_1 and c_2, respectively. Assume that the functions $u \in \mathcal{B}_{c_1}^\varphi$ and $v \in \mathcal{B}_{c_2}^\varphi$ are such that $\boldsymbol{\varphi}_{c_1}(u) = \boldsymbol{\varphi}_{c_2}(v) \in \mathcal{F}_{c_1}^\varphi \cap \mathcal{F}_{c_2}^\varphi$. Then we can write

$$v = c_1 - c_2 + u - (\psi_1(u) - \psi_2(v))u_0.$$

Since the function v is in $B_{c_2}^\varphi$, if we multiply this equation by $\boldsymbol{\varphi}'(c_2)$ and integrate with respect to the measure μ, we obtain

$$0 = \mathbb{E}[(c_1 - c_2 + u)\boldsymbol{\varphi}'(c_2)] - (\psi_1(u) - \psi_2(v))\,\mathbb{E}[u_0\boldsymbol{\varphi}'(c_2)].$$

Thus the transition map $\boldsymbol{\varphi}_{c_2}^{-1} \circ \boldsymbol{\varphi}_{c_1}$ can be expressed as

$$\boldsymbol{\varphi}_{c_2}^{-1} \circ \boldsymbol{\varphi}_{c_1}(w) = c_1 - c_2 + w - \frac{\mathbb{E}[(c_1 - c_2 + w)\boldsymbol{\varphi}'(c_2)]}{\mathbb{E}[u_0\boldsymbol{\varphi}'(c_2)]}u_0, \tag{7}$$

for every $w \in \boldsymbol{\varphi}_{c_1}^{-1}(\mathcal{F}_{c_1}^\varphi \cap \mathcal{F}_{c_2}^\varphi)$. Clearly, this transition map will be of class C^∞ if we show that the functions w and $c_1 - c_2$ are in $L_{c_2}^\varphi$, and the spaces $L_{c_1}^\varphi$ and $L_{c_2}^\varphi$ have equivalent norms. It is not hard to verify that if two Musielak–Orlicz spaces are equal as sets, then their norms are equivalent (see [9, Theorem 8.5]). We make use of the following:

Proposition 4. *Assume that the measurable functions $\widetilde{c}, c \colon T \to \mathbb{R}$ satisfy $\mathbb{E}[\varphi(t, \widetilde{c}(t))] < \infty$ and $\mathbb{E}[\varphi(t, c(t))] < \infty$. Then $L_{\widetilde{c}}^\varphi \subseteq L_c^\varphi$ if and only if $\widetilde{c} - c \in L_c^\varphi$.*

Proof. Suppose that $\widetilde{c} - c$ is not in L_c^φ. Let $A = \{t \in T : \widetilde{c}(t) < c(t)\}$. For $\lambda \in [0, 1]$, we have

$$\begin{aligned}\mathbb{E}[\boldsymbol{\varphi}(c + \lambda(\widetilde{c} - c))] &= \mathbb{E}[\boldsymbol{\varphi}(c + \lambda(\widetilde{c} - c))\mathbf{1}_{T\setminus A}] + \mathbb{E}[\boldsymbol{\varphi}(c + \lambda(\widetilde{c} - c))\mathbf{1}_A] \\ &\leq \mathbb{E}[\boldsymbol{\varphi}(c + (\widetilde{c} - c))\mathbf{1}_{T\setminus A}] + \mathbb{E}[\boldsymbol{\varphi}(c)\mathbf{1}_A] \\ &\leq \mathbb{E}[\boldsymbol{\varphi}(\widetilde{c})] + \mathbb{E}[\boldsymbol{\varphi}(c)] < \infty.\end{aligned}$$

Since $\widetilde{c} - c \notin L_c^\varphi$, for any $\lambda > 0$, there holds $\mathbb{E}[\boldsymbol{\varphi}(c - \lambda(\widetilde{c} - c))] = \infty$. From

$$\begin{aligned}\mathbb{E}[\boldsymbol{\varphi}(c - \lambda(\widetilde{c} - c))] &= \mathbb{E}[\boldsymbol{\varphi}(c - \lambda(\widetilde{c} - c))\mathbf{1}_{T\setminus A}] + \mathbb{E}[\boldsymbol{\varphi}(c - \lambda(\widetilde{c} - c))\mathbf{1}_A] \\ &\leq \mathbb{E}[\boldsymbol{\varphi}(c + \lambda(c - \widetilde{c}))\mathbf{1}_A],\end{aligned}$$

we obtain that $(c - \widetilde{c})\mathbf{1}_A$ does not belong to L^{φ}_{c}. Clearly, $(c - \widetilde{c})\mathbf{1}_A \in L^{\varphi}_{\widetilde{c}}$. Consequently, $L^{\varphi}_{\widetilde{c}}$ is not contained in L^{φ}_{c}.

Conversely, assume $\widetilde{c} - c \in L^{\varphi}_{c}$. Let w be any function in $L^{\varphi}_{\widetilde{c}}$. We can find $\varepsilon > 0$ such that $\mathbb{E}[\boldsymbol{\varphi}(\widetilde{c} + \lambda w)] < \infty$, for every $\lambda \in (-\varepsilon, \varepsilon)$. Consider the convex function

$$g(\alpha, \lambda) = \mathbb{E}[\boldsymbol{\varphi}(c + \alpha(\widetilde{c} - c) + \lambda w)].$$

This function is finite for $\lambda = 0$ and α in the interval $(-\eta, 1]$, for some $\eta > 0$. Moreover, $g(1, \lambda)$ is finite for every $\lambda \in (-\varepsilon, \varepsilon)$. By the convexity of g, we have that g is finite in the convex hull of the set $1 \times (-\varepsilon, \varepsilon) \cup (-\eta, 1] \times 0$. We obtain that $g(0, \lambda)$ is finite for every λ in some neighborhood of 0. Consequently, $w \in L^{\varphi}_{c}$. Since $w \in L^{\varphi}_{c}$ is arbitrary, the inclusion $L^{\varphi}_{\widetilde{c}} \subseteq L^{\varphi}_{c}$ follows. □

Lemma 5. *If the function u is in $\mathcal{K}^{\varphi}_{c}$ and we denote $\widetilde{c} = c + u - \psi(u)u_0$, then the spaces L^{φ}_{c} and $L^{\varphi}_{\widetilde{c}}$ are equal as sets.*

Proof. The inclusion $L^{\varphi}_{\widetilde{c}} \subseteq L^{\varphi}_{c}$ follows from Proposition 4. Since $u \in \mathcal{K}^{\varphi}_{c}$, we have

$$\mathbb{E}[\boldsymbol{\varphi}(\widetilde{c} + \lambda u)] \le \mathbb{E}[\boldsymbol{\varphi}(c + (1 + \lambda)u)] < \infty,$$

for every λ in a neighborhood of 0. Thus $c - \widetilde{c} = -u + \psi(u)u_0$ belongs to $L^{\varphi}_{\widetilde{c}}$. From Proposition 4, we obtain $L^{\varphi}_{\widetilde{c}} \subseteq L^{\varphi}_{c}$. □

By Lemma 5, if we denote $c_1 + u - \psi_1(u)u_0 = \widetilde{c} = c_2 + v - \psi_2(v)u_0$, we have that the spaces $L^{\varphi}_{c_1}$, $L^{\varphi}_{\widetilde{c}}$ and $L^{\varphi}_{c_2}$ are equal as sets. In (7), the function w is in $L^{\varphi}_{c_2}$ and consequently $c_1 - c_2$ is in $L^{\varphi}_{c_2}$. Therefore, the transition map $\boldsymbol{\varphi}^{-1}_{c_2} \circ \boldsymbol{\varphi}_{c_1}$ is of class C^{∞}.

Since $\boldsymbol{\varphi}^{-1}_{c_2} \circ \boldsymbol{\varphi}_{c_1}$ is of class C^{∞}, the set $\boldsymbol{\varphi}^{-1}_{c_1}(\mathcal{F}^{\varphi}_{c_1} \cap \mathcal{F}^{\varphi}_{c_2})$ is open $B^{\varphi}_{c_1}$. The φ-families $\mathcal{F}^{\varphi}_{c}$ are maximal in the sense that if two φ-families $\mathcal{F}^{\varphi}_{c_1}$ and $\mathcal{F}^{\varphi}_{c_2}$ have non-empty intersection, then they coincide.

Lemma 6. *For a function u in $\mathcal{B}^{\varphi}_{c}$, denote $\widetilde{c} = c + u - \psi(u)u_0$. Then $\mathcal{F}^{\varphi}_{c} = \mathcal{F}^{\varphi}_{\widetilde{c}}$.*

Proof. Let v be a function in $\mathcal{B}^{\varphi}_{c}$. Then there exists $\varepsilon > 0$ such that, for every $\lambda \in (-\varepsilon, 1 + \varepsilon)$, the function $\boldsymbol{\varphi}(c + \lambda v + (1 - \lambda)u)$ is μ-integrable. Consequently, $\varphi(\widetilde{c} + \lambda(v - u))$ is μ-integrable for all $\lambda \in (-\varepsilon, 1 + \varepsilon)$. Thus the difference $v - u$ is in $\mathcal{K}^{\varphi}_{\widetilde{c}}$ and

$$w = v - u - \frac{\mathbb{E}[(v - u)\boldsymbol{\varphi}'(\widetilde{c})]}{\mathbb{E}[u_0 \boldsymbol{\varphi}'(\widetilde{c})]} u_0 \tag{8}$$

belongs to $\mathcal{B}^{\varphi}_{\widetilde{c}}$. Let $\widetilde{\psi}\colon \mathcal{B}^{\varphi}_{\widetilde{c}} \to [0, \infty)$ be the normalizing function associated to $\widetilde{c}$. Then the probability density $\boldsymbol{\varphi}(\widetilde{c} + w - \widetilde{\psi}(w)u_0)$ is in $\mathcal{F}^{\varphi}_{\widetilde{c}}$. This probability density can be expressed as $\boldsymbol{\varphi}(c + v - ku_0)$ for a constant k. According to Proposition 3, there exists a unique $\psi(u) \in \mathbb{R}$ such that the probability density $\boldsymbol{\varphi}(c + v - \psi(v)u_0)$ is in $\mathcal{F}^{\varphi}_{c}$. Therefore, $\mathcal{F}^{\varphi}_{c} \subseteq \mathcal{F}^{\varphi}_{\widetilde{c}}$.

Using the same arguments as in the previous paragraph, we obtain that $c = \widetilde{c} + w - \widetilde{\psi}(w)u_0$, where the function $w \in \mathcal{B}^{\varphi}_{\widetilde{c}}$ is given in (8) with $v = 0$. Thus $\mathcal{F}^{\varphi}_{\widetilde{c}} \subseteq \mathcal{F}^{\varphi}_{c}$. □

By Lemma 6, if we denote $c_1 + u - \psi_1(u)u_0 = \widetilde{c} = c_2 + v - \psi_2(v)u_0$, then we have the equality $\mathcal{F}^{\varphi}_{c_1} = \mathcal{F}^{\varphi}_{\widetilde{c}} = \mathcal{F}^{\varphi}_{c_2}$.

The results obtained in these lemmas are summarized in the next Proposition.

Proposition 7. *Let $c_1, c_2 \colon T \to \mathbb{R}$ be measurable functions such that the probability densities $p_1 = \boldsymbol{\varphi}(c_1)$ and $p_2 = \boldsymbol{\varphi}(c_2)$ are in $\mathcal{P}_\mu$. Suppose $\mathcal{F}^{\varphi}_{c_1} \cap \mathcal{F}^{\varphi}_{c_2} \neq \emptyset$. Then the Musielak–Orlicz spaces $L^{\varphi}_{c_1}$ and $L^{\varphi}_{c_2}$ are equal as sets, and have equivalent norms. Moreover, $\mathcal{F}^{\varphi}_{c_1} = \mathcal{F}^{\varphi}_{c_2}$.*

The collection $\{(\mathcal{B}^{\varphi}_c, \boldsymbol{\varphi}_c)\}_{\boldsymbol{\varphi}(c)\in\mathcal{P}_\mu}$ satisfies (bm1)–(bm2), equipping $\mathcal{P}_\mu$ with a C^∞-differentiable structure.

5. Divergence

In this section we define the divergence between two probability distributions. The entities found in Information Geometry [5, 6], like the Fisher information, connections, geodesics, etc., are all derived from the divergence taken in the considered family. The divergence we will found is the Bregman divergence [15] associated to the normalizing function $\psi \colon \mathcal{K}^{\varphi}_c \to [0, \infty)$. We show that our divergence does not depend on the parametrization of the φ-family $\mathcal{F}^{\varphi}_c$.

Let S be a convex subset of a Banach space X. Given a convex function $f \colon S \to \mathbb{R}$, the *Bregman divergence* $B_f \colon S \times S \to [0, \infty)$ is defined as

$$B_f(y, x) = f(y) - f(x) - \partial_+ f(x)(y - x),$$

for all $x, y \in S$, where $\partial_+ f(x)(h) = \lim_{t\downarrow 0}(f(x + th) - f(x))/t$ denotes the *right-directional derivative* of f at x in the direction of h. The right-directional derivative $\partial_+ f(x)(h)$ exists and defines a sublinear functional. If the function f is strictly convex, the divergence satisfies $B_f(y, x) = 0$ if and only if $x = y$.

Let X and Y be Banach spaces, and $U \subseteq X$ be an open set. A function $f \colon U \to Y$ is said to be *Gâteaux-differentiable* at $x_0 \in U$ if there exists a bounded linear map $A \colon X \to Y$ such that

$$\lim_{t\to 0} \frac{1}{t} \|f(x_0 + th) - f(x_0) - Ah\| = 0,$$

for every $h \in X$. The *Gâteaux derivative* of f at x_0 is denoted by $A = \partial f(x_0)$. If the limit above can be taken uniformly for every $h \in X$ such that $\|h\| \le 1$, then the function f is said to be *Fréchet-differentiable* at x_0. The *Fréchet derivative* of f at x_0 is denoted by $A = Df(x_0)$.

Now we verify that $\psi \colon \mathcal{K}^{\varphi}_c \to \mathbb{R}$ is a convex function. Take any $u, v \in \mathcal{K}^{\varphi}_c$ such that $u \neq v$. Clearly, the function $\lambda u + (1-\lambda)v$ is in $\mathcal{K}^{\varphi}_c$, for any $\lambda \in (0, 1)$. By the convexity of $\varphi(t, \cdot)$, we can write

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{\varphi}(c + \lambda u + (1-\lambda)v - \lambda\psi(u)u_0 - (1-\lambda)\psi(v)u_0)] \\
&\qquad \le \lambda\, \mathbb{E}[\boldsymbol{\varphi}(c + u - \psi(u)u_0)] + (1-\lambda)\, \mathbb{E}[\boldsymbol{\varphi}(c + v - \psi(v)u_0)] = 1.
\end{aligned}$$

Since $\boldsymbol{\varphi}(c + \lambda u + (1-\lambda)v - \psi(\lambda u + (1-\lambda)v)u_0)$ has μ-integral equal to 1, we can conclude that the following inequality holds:

$$\psi(\lambda u + (1-\lambda)v) \leq \lambda\psi(u) + (1-\lambda)\psi(v).$$

So we can define the Bregman divergence B_ψ from to the normalizing function ψ.

The Bregman divergence $B_\psi \colon \mathcal{B}_c^\varphi \times \mathcal{B}_c^\varphi \to [0,\infty)$ associated to the normalizing function $\psi \colon \mathcal{B}_c^\varphi \to [0,\infty)$ is given by

$$B_\psi(v,u) = \psi(v) - \psi(u) - \partial_+\psi(u)(v-u).$$

Then we define the divergence $D_\psi \colon \mathcal{B}_c^\varphi \times \mathcal{B}_c^\varphi \to [0,\infty)$ related to the φ-family $\mathcal{F}_c^\varphi$ as

$$D_\psi(u,v) = B_\psi(v,u).$$

The entries of B_ψ are inverted in order that D_ψ corresponds in some way to the *Kullback–Leibler divergence* $D_{\mathrm{KL}}(p,q) = \mathbb{E}[p\log(\frac{p}{q})]$. Assuming that $\varphi(t,\cdot)$ is continuously differentiable (or strictly convex), we will find an expression for $\partial\psi(u)$.

Lemma 8. *Assume that $\varphi(t,\cdot)$ is continuously differentiable. For any $u \in \mathcal{K}_c^\varphi$, the linear functional $f_u \colon L_c^\varphi \to \mathbb{R}$ given by $f_u(v) = \mathbb{E}[v\boldsymbol{\varphi}'(c+u)]$ is bounded.*

Proof. Every function $v \in L_c^\varphi$ with norm $\|v\|_{\Phi,0} \leq 1$ satisfies $I_\Phi(v) \leq \|u\|_{\Phi,0}$. Then we obtain

$$\mathbb{E}[\boldsymbol{\varphi}(c+|v|)] = I_\Phi(v) + \mathbb{E}[\boldsymbol{\varphi}(c)] \leq 2.$$

Since $u \in \mathcal{K}_c^\varphi$, we can find $\lambda \in (0,1)$ such that $\mathbb{E}[\boldsymbol{\varphi}(c + \frac{1}{\lambda}u)] < \infty$. We can write

$$\begin{aligned}(1-\lambda)\,\mathbb{E}[|v|\boldsymbol{\varphi}'(c+u)] &\leq \mathbb{E}[\boldsymbol{\varphi}(c+u+(1-\lambda)|v|)] - \mathbb{E}[\boldsymbol{\varphi}(c+u)] \\ &= \mathbb{E}[\boldsymbol{\varphi}(\lambda(c+\tfrac{1}{\lambda}u) + (1-\lambda)(c+|v|))] - \mathbb{E}[\boldsymbol{\varphi}(c+u)] \\ &\leq \lambda\,\mathbb{E}[\boldsymbol{\varphi}(c+\tfrac{1}{\lambda}u)] + (1-\lambda)\,\mathbb{E}[\boldsymbol{\varphi}(c+|v|)] - \mathbb{E}[\boldsymbol{\varphi}(c+u)].\end{aligned}$$

Thus the absolute value of $f_u(v) = \mathbb{E}[v\boldsymbol{\varphi}'(c+u)]$ is bounded by some constant for $\|v\|_{\Phi,0} \leq 1$. □

Lemma 9. *Assume that $\varphi(t,\cdot)$ is continuously differentiable. Then the normalizing function $\psi \colon \mathcal{K}_c^\varphi \to \mathbb{R}$ is Gâteaux-differentiable and*

$$\partial\psi(u)v = \frac{\mathbb{E}[v\boldsymbol{\varphi}'(c+u-\psi(u)u_0)]}{\mathbb{E}[u_0\boldsymbol{\varphi}'(c+u-\psi(u)u_0)]}. \tag{9}$$

Proof. According to Lemma 8, the expression in (9) defines a bounded linear functional. Fix functions $u \in \mathcal{K}_c^\varphi$ and $v \in L_c^\varphi$. In virtue of Proposition 4, we can find $\varepsilon > 0$ such that $\mathbb{E}[\boldsymbol{\varphi}(c+u+\lambda|v|)] < \infty$, for every $\lambda \in [-\varepsilon,\varepsilon]$. Define

$$g(\lambda,k) = \mathbb{E}[\boldsymbol{\varphi}(c+u+\lambda v - ku_0)],$$

for any $\lambda \in (-\varepsilon, \varepsilon)$ and $k \geq 0$. Since $\mathcal{K}_c^{\varphi}$ is open, there exist a sufficiently small $\alpha_0 > 0$ such that $u + \lambda v + \alpha|v|$ is in $\mathcal{K}_c^{\varphi}$ for all $\alpha \in [-\alpha_0, \alpha_0]$. We can write

$$\frac{g(\lambda+\alpha,k)-g(\lambda,k)}{\alpha} = \mathbb{E}\Big[\frac{1}{\alpha}\{\boldsymbol{\varphi}(c+u+(\lambda+\alpha)v-ku_0)-\boldsymbol{\varphi}(c+u+\lambda v-ku_0)\}\Big].$$

The function in the expectation above is dominated by the μ-integrable function $\frac{1}{\alpha_0}\{\boldsymbol{\varphi}(c+u+\lambda v+\alpha_0|v|-ku_0)-\boldsymbol{\varphi}(c+u+\lambda v-ku_0)\}$. By the Dominated Convergence Theorem,

$$\mathbb{E}\Big[\frac{1}{\alpha}\{\boldsymbol{\varphi}(c+u+(\lambda+\alpha)v-ku_0)-\boldsymbol{\varphi}(c+u+\lambda v-ku_0)\}\Big] \\ \to \mathbb{E}[v\boldsymbol{\varphi}'(c+u+\lambda v-ku_0)], \qquad \text{as } \alpha \to 0,$$

and, consequently,

$$\frac{\partial g}{\partial \lambda}(\lambda,k) = \mathbb{E}[v\boldsymbol{\varphi}'(c+u+\lambda v-ku_0)].$$

Since $v\boldsymbol{\varphi}'(c+u+\lambda v-ku_0)$ is dominated by the μ-integrable function $|v|\boldsymbol{\varphi}'(c+u+\varepsilon|v|-ku_0)$, we obtain for any sequence $\lambda_n \to \lambda$,

$$\mathbb{E}[v\boldsymbol{\varphi}'(c+u+\lambda_n v-ku_0)] \to \mathbb{E}[v\boldsymbol{\varphi}'(c+u+\lambda v-ku_0)], \quad \text{as } n \to \infty.$$

Thus $\frac{\partial g}{\partial \lambda}(\lambda,k)$ is continuous with respect to λ. Analogously, it can be shown that

$$\frac{\partial g}{\partial k}(\lambda,k) = -\,\mathbb{E}[u_0\boldsymbol{\varphi}'(c+u+\lambda v-ku_0)],$$

and $\frac{\partial g}{\partial k}(\lambda,k)$ is continuous with respect to k. The equality $g(\lambda,k(\lambda)) = \mathbb{E}[\boldsymbol{\varphi}(c+u+\lambda v-k(\lambda)u_0)] = 1$ defines $k(\lambda) = \psi(u+\lambda v)$ as an implicit function of λ. Notice that $\frac{\partial g(0,k)}{\partial k} < 0$. By the Implicit Function Theorem, the function $k(\lambda) = \psi(u+\lambda v)$ is continuously differentiable in a neighborhood of 0, and has derivative

$$\frac{\partial k}{\partial \lambda}(0) = -\frac{(\partial g/\partial \lambda)(0,k(0))}{(\partial g/\partial k)(0,k(0))}.$$

Consequently,

$$\partial\psi(u)(v) = \frac{\partial\psi(u+\lambda v)}{\partial\lambda}(0) = \frac{\mathbb{E}[v\boldsymbol{\varphi}'(c+u-\psi(u)u_0)]}{\mathbb{E}[u_0\boldsymbol{\varphi}'(c+u-\psi(u)u_0)]}.$$

Thus the expression in (9) is the Gâteaux-derivative of ψ. □

Lemma 10. *Assume that $\varphi(t,\cdot)$ is continuously differentiable. Then the divergence D_ψ does not depend on the parametrization of $\mathcal{F}_c^{\varphi}$.*

Proof. For any $w \in \mathcal{B}_c^{\varphi}$, we denote $\widetilde{c} = c + w - \psi(w)u_0$. Given $u, v \in \mathcal{B}_c^{\varphi}$, select $\widetilde{u}, \widetilde{v} \in \mathcal{B}_{\widetilde{c}}^{\varphi}$ such that $\boldsymbol{\varphi}_{\widetilde{c}}(\widetilde{u}) = \boldsymbol{\varphi}_c(u)$ and $\boldsymbol{\varphi}_{\widetilde{c}}(\widetilde{v}) = \boldsymbol{\varphi}_c(v)$. Let $\widetilde{\psi}\colon \mathcal{B}_{\widetilde{c}}^{\varphi} \to [0,\infty)$ be the normalizing function associated to $\widetilde{c}$. These definitions provide

$$\widetilde{c} + \widetilde{u} - \widetilde{\psi}(\widetilde{u})u_0 = c + u - \psi(u)u_0,$$

and

$$\widetilde{c} + \widetilde{v} - \widetilde{\psi}(\widetilde{v})u_0 = c + v - \psi(v)u_0.$$

Subtracting these equations, we obtain

$$[-\widetilde{\psi}(\widetilde{v}) + \widetilde{\psi}(\widetilde{u})]u_0 + (\widetilde{v} - \widetilde{u}) = [-\psi(v) + \psi(u)]u_0 + (v - u)$$

and, consequently,

$$\begin{aligned}\widetilde{\psi}(\widetilde{v}) - \widetilde{\psi}(\widetilde{u}) - \frac{\mathbb{E}[(\widetilde{v} - \widetilde{u})\boldsymbol{\varphi}'(\widetilde{c} + \widetilde{u} - \widetilde{\psi}(\widetilde{u})u_0)]}{\mathbb{E}[u_0\boldsymbol{\varphi}'(\widetilde{c} + \widetilde{u} - \widetilde{\psi}(\widetilde{u})u_0)]}& \\ = \psi(v) - \psi(u) - \frac{\mathbb{E}[(v-u)\boldsymbol{\varphi}'(c + u - \psi(u)u_0)]}{\mathbb{E}[u_0\boldsymbol{\varphi}'(c + u - \psi(u)u_0)]}&.\end{aligned}$$

Therefore, $D_{\widetilde{\psi}}(\widetilde{u}, \widetilde{v}) = D_\psi(u, v)$. □

Let $p = \boldsymbol{\varphi}_c(u)$ and $q = \boldsymbol{\varphi}_c(v)$, for $u, v \in \mathcal{B}_c^\varphi$. We denote the divergence between the probability densities p and q by

$$D(p \parallel q) = D_\psi(u, v).$$

According to Lemma 10, $D(p \parallel q)$ is well-defined if p and q are in the same φ-family. We will find an expression for $D(p \parallel q)$ where p and q are given explicitly. For $u = 0$, we have $D(p \parallel q) = D_\psi(0, v) = \psi(v)$, and then

$$D(p \parallel q) = \frac{\mathbb{E}[(-v + \psi(v)u_0)\boldsymbol{\varphi}'(c)]}{\mathbb{E}[u_0\boldsymbol{\varphi}'(c)]}.$$

Therefore, the divergence between probability densities p and q in the same φ-family can be expressed as

$$D(p \parallel q) = \frac{\mathbb{E}\left[\dfrac{\boldsymbol{\varphi}^{-1}(p) - \boldsymbol{\varphi}^{-1}(q)}{(\boldsymbol{\varphi}^{-1})'(p)}\right]}{\mathbb{E}\left[\dfrac{u_0}{(\boldsymbol{\varphi}^{-1})'(p)}\right]}. \tag{10}$$

Clearly, the expectation in (10) may not be defined if p and q are not in the same φ-family. We extend the divergence in (10) by setting $D(p \parallel q) = \infty$ if p and q are not in the same φ-family. With this extension, the divergence is denoted by D_φ and is called the *φ-divergence*. By the strict convexity of $\varphi(t, \cdot)$, we have the inequality $\varphi^{-1}(t, u) - \varphi^{-1}(t, v) \geq (\varphi^{-1})'(t, u)(u - v)$ for any $u, v > 0$, with equality if and only if $u = v$. Hence D_φ is always non-negative, and $D_\varphi(p \parallel q)$ is equal to zero if and only if $p = q$.

Example 11. With the variable κ-exponential $\exp_\kappa(t, u) = \exp_{\kappa(t)}(u)$ in the place of $\varphi(t, u)$, whose inverse $\varphi^{-1}(t, u)$ is the variable κ-logarithm $\ln_\kappa(t, u) = \ln_{\kappa(t)}(u)$, we rewrite (10) as

$$D(p \parallel q) = \frac{\mathbb{E}\left[\dfrac{\mathbf{ln}_{\boldsymbol{\kappa}}(p) - \mathbf{ln}_{\boldsymbol{\kappa}}(q)}{\mathbf{ln}'_{\boldsymbol{\kappa}}(p)}\right]}{\mathbb{E}\left[\dfrac{u_0}{\mathbf{ln}'_{\boldsymbol{\kappa}}(p)}\right]}, \tag{11}$$

where $\mathbf{ln}_{\boldsymbol{\kappa}}(p)$ denotes $\ln_{\kappa(t)}(p(t))$. Since the κ-logarithm $\ln_\kappa(u) = \frac{u^\kappa - u^{-\kappa}}{2\kappa}$ has derivative $\ln'_\kappa(u) = \frac{1}{u}\frac{u^\kappa + u^{-\kappa}}{2}$, the numerator and denominator in (11) result in

$$\mathbb{E}\left[\frac{\mathbf{ln}_{\boldsymbol{\kappa}}(p) - \mathbf{ln}_{\boldsymbol{\kappa}}(q)}{\mathbf{ln}'_{\boldsymbol{\kappa}}(p)}\right] = \mathbb{E}\left[\frac{\dfrac{p^\kappa - p^{-\kappa}}{2\kappa} - \dfrac{q^\kappa - q^{-\kappa}}{2\kappa}}{\dfrac{1}{p}\dfrac{p^\kappa + p^{-\kappa}}{2}}\right] = \frac{1}{\kappa}\,\mathbb{E}_p\left[\frac{p^\kappa - p^{-\kappa}}{p^\kappa + p^{-\kappa}} - \frac{q^\kappa - q^{-\kappa}}{p^\kappa + p^{-\kappa}}\right]$$

and

$$\mathbb{E}\left[\frac{u_0}{\mathbf{ln}'_{\boldsymbol{\kappa}}(p)}\right] = \mathbb{E}_p\left[\frac{2u_0}{p^\kappa + p^{-\kappa}}\right],$$

respectively. Thus (11) can be rewritten as

$$D_\kappa(p \,\|\, q) = \frac{1}{\kappa}\frac{\mathbb{E}_p\left[\dfrac{p^\kappa - p^{-\kappa}}{p^\kappa + p^{-\kappa}} - \dfrac{q^\kappa - q^{-\kappa}}{p^\kappa + p^{-\kappa}}\right]}{\mathbb{E}_p\left[\dfrac{2u_0}{p^\kappa + p^{-\kappa}}\right]},$$

which we called the κ-*divergence*.

References

[1] G. Pistone, C. Sempi, An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one, Ann. Statist. 23 (5) (1995) 1543–1561. `doi:10.1214/aos/1176324311`.

[2] G. Pistone, M. P. Rogantin, The exponential statistical manifold: mean parameters, orthogonality and space transformations, Bernoulli 5 (4) (1999) 721–760. `doi:10.2307/3318699`.

[3] A. Cena, G. Pistone, Exponential statistical manifold, Ann. Inst. Statist. Math. 59 (1) (2007) 27–56. `doi:10.1007/s10463-006-0096-y`.

[4] M. R. Grasselli, Dual connections in nonparametric classical information geometry, Ann. Inst. Statist. Math. 62 (5) (2010) 873–896. `doi:10.1007/s10463-008-0191-3`.

[5] S.-i. Amari, H. Nagaoka, Methods of information geometry, Vol. 191 of Translations of Mathematical Monographs, American Mathematical Society, Providence, RI, 2000, translated from the 1993 Japanese original by Daishi Harada.

[6] M. K. Murray, J. W. Rice, Differential geometry and statistics, Vol. 48 of Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1993.

[7] P. Gibilisco, E. Riccomagno, M. P. Rogantin, H. P. Wynn (Eds.), Algebraic and geometric methods in statistics, Cambridge University Press, Cambridge, 2010.

[8] D. Petz, Quantum information theory and quantum statistics, Theoretical and Mathematical Physics, Springer-Verlag, Berlin, 2008.

[9] J. Musielak, Orlicz spaces and modular spaces, Vol. 1034 of Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1983.

[10] M. A. Krasnosel′skiĭ, J. B. Rutickiĭ, Convex functions and Orlicz spaces, Translated from the first Russian edition by Leo F. Boron, P. Noordhoff Ltd., Groningen, 1961.

[11] M. M. Rao, Z. D. Ren, Theory of Orlicz spaces, Vol. 146 of Monographs and Textbooks in Pure and Applied Mathematics, Marcel Dekker Inc., New York, 1991.

[12] G. Kaniadakis, M. Lissia, A. M. Scarfone, Deformed logarithms and entropies, Phys. A 340 (1-3) (2004) 41–49, news and expectations in thermostatistics. doi:10.1016/j.physa.2004.03.075.

[13] J. Naudts, Estimators, escort probabilities, and ϕ-exponential families in statistical physics, JIPAM. J. Inequal. Pure Appl. Math. 5 (4) (2004) Article 102, 15 pp. (electronic).

[14] J. Naudts, Generalised exponential families and associated entropy functions, Entropy 10 (3) (2008) 131–149. doi:10.3390/entropy-e10030131.

[15] L. M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR Computational Mathematics and Mathematical Physics 7 (3) (1967) 200–217. doi:10.1016/0041-5553(67)90040-7.